

# Unfolding Latent Tree Structures using 4th Order Tensors

Mariya Ishteva, Haesun Park, Le Song

*College of Computing, Georgia Institute of Technology*  
 {mishteva,hpark,lsong}@cc.gatech.edu

## Abstract

Discovering the latent structure from many observed variables is an important yet challenging learning task. Existing approaches for discovering latent structures often require the unknown number of hidden states as an input. In this paper, we propose a quartet based approach which is *agnostic* to this number. The key contribution is a novel rank characterization of the tensor associated with the marginal distribution of a quartet. This characterization allows us to design a *nuclear norm* based test for resolving quartet relations. We then use the quartet test as a subroutine in a divide-and-conquer algorithm for recovering the latent tree structure. Under mild conditions, the algorithm is consistent and its error probability decays exponentially with increasing sample size. We demonstrate that the proposed approach compares favorably to alternatives. In a real world stock dataset, it also discovers meaningful groupings of variables, and produces a model that fits the data better.

## 1 Introduction

Discovering the latent structure from many observed variables is an important yet challenging learning task. The discovered structures can help better understand the domain and lead to potentially better predictive models. Many local search heuristics based on maximum parsimony and maximum likelihood methods have been proposed to address this problem (Semple & Steel, 2003; Zhang, 2004; Heller & Ghahramani, 2005; Teh et al., 2008; Harmeling & Williams, 2010). Their common drawback is that it is difficult to provide consistency guarantees. Furthermore, the number of hidden states often needs to be determined before the structure learning. Or cross-validations are needed to determine the hidden states, which can be very time consuming to run.

Efficient algorithms with provable performance guarantees have been explored in the phylogenetic tree reconstruction community. One popular algorithm is the neighbor-joining (NJ) algorithm (Saitou & Nei, 1987), where pairs of variables are joined recursively according to a certain distance measure. The NJ algorithm is consistent when the distance measure satisfies the path additive property (Mihaescu et al., 2009). For discrete random variables, the additive distance is defined using the determinant of the joint probability table of a pair of variables (Lake, 1994). However, this definition only applies to the cases where the observed variables and latent variables have the same number of states. When the latent variables represent simpler factors with smaller number of states, the NJ algorithm can perform poorly.

Another family of provably consistent reconstruction methods is the quartet-based methods (Semple & Steel, 2003; Erdős et al., 1999). These methods first resolve a set of latent relations

for quadruples of observed variables (quartets), and subsequently, stitch them together to form a latent tree. A good quartet test plays an essential role in these methods, as it is called repeatedly by the stitching algorithms. Recently, (Anandkumar et al., 2011) proposed a quartet test using the leading  $k$  singular values of the joint probability table, where  $k$  is the number of hidden states. This new approach allows  $k$  to be different from the number of the observed states. However, it still requires  $k$  to be given in advance.

Our goal is to design a latent structure discovery algorithm which is *agnostic* to the number of hidden states, since in practice we rarely know this number. The proposed approach is quartet based, where the quartet relations are resolved based on rank properties of 4th order tensors associated with the joint probability tables of quartets. The key insight is that rank properties of the tensor reveal the latent structure behind a quartet. Similar observations have been reported in the phylogenetic community (Eriksson, 2005; Allman & Rhodes, 2006), but they are concerned about the cases where the number of hidden states is larger or equal to the number of observed states. We focus instead on the cases where the number of hidden states is smaller, representing simpler factors. Furthermore, if the joint probability tensor is only approximately given (due to sampling noise) the main rank condition has to be modified. In Allman & Rhodes (2006) such condition is missing and in Eriksson (2005) the condition is heuristically translated to the distance of a matrix to its best rank- $k$  approximation. In contrast, we propose a novel nuclear norm relaxation of the rank condition, discuss its advantages, and provide recovery conditions and finite sample guarantees. Our quartet test is easy to compute since it only involves singular value decomposition of unfolded 4th order tensors.

Using the proposed quartet test as a subroutine, the latent tree structure can be recovered in a divide-and-conquer fashion (Pearl & Tarsi, 1986). For  $d$  observed variables, the computational complexity of the algorithm is  $O(d \log d)$ , making it scalable to large problems. Under mild conditions, the tree construction algorithm using our quartet test is consistent and stable to estimate given a finite number of samples. In simulations, we compared to alternatives in terms of resolving quartet relations and building the entire latent trees. The proposed approach is among the best performing ones while being agnostic to the number of hidden states  $k$ . The latter is an important improvement, since cross validation for finding  $k$  is expensive while leading to similar final results. We also applied the new approach to a stock dataset, where it discovered meaningful grouping of stocks according to industrial sectors, and led a latent variable model that fits the data better than the competitors.

## 2 Latent Tree Graphical Models

In this paper, we focus on discrete latent variable models where the conditional independence structures are specified by trees. We assume that the  $d$  observed variables,  $\mathcal{O} = \{X_1, \dots, X_d\}$ , are leaves of the tree and that they all have the same number of states,  $n$ . We also assume the  $d_h$  hidden variables,  $\mathcal{H} = \{X_{d+1}, \dots, X_{d+d_h}\}$ , have the same<sup>1</sup>, *but unknown*, number of states,  $k$ , ( $k \leq n$ ). Furthermore, we use uppercase letters to denote random variables (*e.g.*,  $X_i$ ) and lowercase letters their instantiations (*e.g.*,  $x_i$ ).

**Factorization of distribution.** The joint distribution of all variables,  $\mathcal{X} = \mathcal{O} \cup \mathcal{H}$ , in a latent tree model is a multi-way table (tensor),  $\mathcal{P}$ , with  $d + d_h$  dimensions. Although the tensor

---

<sup>1</sup>Our results are easily generalizable to the case where all hidden variables have different number of states.

has  $O(n^d k^{d_h})$  number of entries, they can be computed from just a polynomial number of parameters due to the latent tree structure. That is  $\mathcal{P}(x_1, \dots, x_{d+d_h}) = \prod_{i=1}^{d+d_h} P(x_i | x_{\pi_i})$  where each  $P(X_i | X_{\pi_i})$  is a conditional probability table (CPT) of a variable  $X_i$  and its parent  $X_{\pi_i}$  in the tree.<sup>2</sup> This factorization leads to a significant saving in terms of tensor representation: we can represent exponential number of entries using just  $O(d_h k^2 + dnk)$  parameters from the CPTs. Throughout the paper, we assume that **(A1)** all CPTs have full column rank,  $k$ .

**Structure learning.** Determining the tree topology  $\mathcal{T}$  is an important and challenging learning problem. The goal is to discover the latent structure based just on samples from observed variables. For simplicity and uniqueness of the tree topology (Pearl, 1988), we assume that **(A2)** every latent variable has *exactly* 3 neighbors.

**Quartet.** A quadruple of observed variables from a latent tree  $\mathcal{T}$  is called a quartet (Figure 1). Under assumption **(A2)**, there are 3 ways to connect a quartet,  $X_1, X_2, X_3, X_4$ , using 2 latent vari-

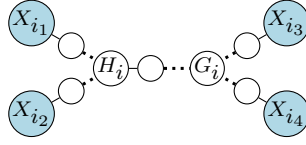


Figure 1: Quartet  $(X_1, X_2, X_3, X_4)$  from a tree.

ables  $H$  and  $G$  (Figure 2). However, only one of the 3 quartet relations is consistent with  $\mathcal{T}$ . The

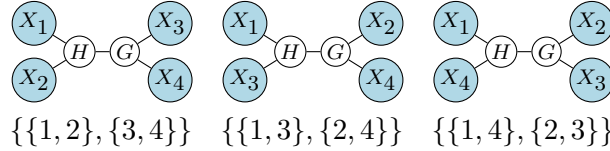


Figure 2: Three fixed ways to connect  $X_1, X_2, X_3, X_4$ , with two latent variables  $H$  and  $G$ .

mapping between quartets and the tree topology  $\mathcal{T}$  is captured in the following theorem (Buneman, 1971):

**Theorem 1.** *The set of all quartet relations  $\mathcal{Q}_{\mathcal{T}}$  is unique to a latent tree  $\mathcal{T}$ , and furthermore,  $\mathcal{T}$  can be recovered from  $\mathcal{Q}_{\mathcal{T}}$  in polynomial time.*

**Quartet-based tree reconstruction.** Motivated by Theorem 1, a family of latent tree recovery algorithms has been designed based on resolving quartet relations. These algorithms first determine one of the 3 ways how 4 variables are connected, and then join together all quartet relations to form a consistent latent tree. For a model with  $d$  observed variables, there are  $O(d^4)$  quartet relations in total (taking all possible combinations of 4 variables). However, we do not necessarily need to resolve all these quartet relations in order to reconstruct the latent tree. A small set of size  $O(d \log d)$  will suffice for the tree recovery, which makes quartet based methods efficient even for problems with large  $d$  (Pearl & Tarsi, 1986; Pearl, 1988). In this paper, we design a new quartet based method. Our main contribution compared to previous approaches is that our method is *agnostic* to the number of hidden states,  $k$ , which is usually unknown in practice.

<sup>2</sup>For a latent tree, we can select a latent node as the root, and re-orient all edges away from it to induce consistent parent-child relations. For the root node  $X_r$ ,  $P(X_r | X_{\pi_r}) = P(X_r)$ .

### 3 Resolving Quartet Relations without Knowing the Number of Hidden States

In this section, we develop a test for resolving the latent relation of a quartet when the number of hidden states is unknown. Our approach makes use of information from the joint probability table of a quartet, which is a 4-way table or 4th order tensor. Suppose that the quartet relation of 4 variables,  $X_1, X_2, X_3$  and  $X_4$ , is  $\{\{1, 2\}, \{3, 4\}\}$ , then the entries in this tensor are specified by

$$\mathcal{P}(x_1, x_2, x_3, x_4) = \sum_{h,g} P(x_1|h)P(x_2|h)P(h,g)P(x_3|g)P(x_4|g). \quad (1)$$

This factorization suggests that there exist some low rank structures in the 4th order tensor. To study the rank properties of  $\mathcal{P}(X_1, X_2, X_3, X_4)$ , we first relate it to the conditional probability tables,  $P(X_1|H)$ ,  $P(X_2|H)$ ,  $P(X_3|G)$ ,  $P(X_4|G)$ , and the joint probability table,  $P(H, G)$  (we abbreviate them as  $P_{1|H}$ ,  $P_{2|H}$ ,  $P_{3|G}$ ,  $P_{4|G}$  and  $P_{HG}$ , respectively). Using tensor algebra, we have

$$\mathcal{P}(X_1, X_2, X_3, X_4) = \langle \mathcal{T}_1, \mathcal{T}_2 \rangle_3,$$

$$\begin{aligned} \text{with } \mathcal{T}_1 &= \mathcal{I}_H \times_1 P_{1|H} \times_2 P_{2|H}, \\ \mathcal{T}_2 &= \mathcal{I}_G \times_1 P_{3|G} \times_2 P_{4|G} \times_3 P_{HG}, \end{aligned}$$

where  $\mathcal{I}_H$  and  $\mathcal{I}_G$  are 3rd order diagonal tensors of size  $k \times k \times k$  with diagonal elements equal to 1. The multiplication  $\times_i$  denotes a tensor-matrix multiplication with respect to the  $i$ -th dimension of the tensor and the rows of the matrix, and  $\langle \cdot, \cdot \rangle_3$  denotes tensor-tensor multiplication along the third dimension of both tensors<sup>3</sup>. This formula can be schematically understood as Figure 3. We

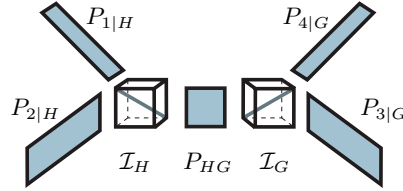


Figure 3: Schematic diagram of the tensor  $\mathcal{P}(X_1, X_2, X_3, X_4)$ .

will start by characterizing the rank properties of  $\mathcal{P}$  and then exploit them to design a quartet test. Although the proposed approach involves unfolding the tensor and subsequent computation at the matrix level, modeling the problem using tensors provides higher level conceptual understanding of the structure of  $\mathcal{P}$ . The novelty of our use of low rank tensors is for latent structure discovery.

#### 3.1 Unfolding the 4th Order Tensor

Now we consider 3 different reshapings  $A$ ,  $B$  and  $C$  of the tensor into matrices (“unfoldings”). These unfoldings contain exactly the same entries as  $\mathcal{P}$  but in different order.  $A$  corresponds to the grouping  $\{\{1, 2\}, \{3, 4\}\}$  of the variables, *i.e.*, the rows of  $A$  correspond to dimensions 1 and 2 of  $\mathcal{P}$ , and its columns to dimensions 3 and 4.  $B$  corresponds to the grouping  $\{\{1, 3\}, \{2, 4\}\}$  and  $C$  - to the grouping  $\{\{1, 4\}, \{2, 3\}\}$ . Using MATLAB’s notation (see appendix, §8 for further explanation),

<sup>3</sup>For formal definitions of tensor notations see appendix, §8.

$$A = \text{reshape}(\mathcal{P}, n^2, n^2); \quad (2)$$

$$B = \text{reshape}(\text{permute}(\mathcal{P}, [1, 3, 2, 4]), n^2, n^2); \quad (3)$$

$$C = \text{reshape}(\text{permute}(\mathcal{P}, [1, 4, 2, 3]), n^2, n^2). \quad (4)$$

Next we present useful characterizations of  $A$ ,  $B$  and  $C$ , which will be essential for understanding their connection with the latent structure of a quartet. The *Kronecker product* of two matrices  $M$  and  $M'$  is denoted as  $M \otimes M'$ , and if they have the same number of columns, their *Khatri-Rao product* (column-wise Kronecker product), is denoted as  $M \odot M'$ . Then (see appendix §9 for proof),

**Lemma 2.** *Assume that  $\{\{1, 2\}, \{3, 4\}\}$  is the correct latent structure. The matrices  $A$ ,  $B$  and  $C$  can be factorized respectively as (see Figure 4(a) and Figure 4(b) for schematic diagrams)*

$$A = (P_{2|H} \odot P_{1|H}) P_{HG} (P_{4|G} \odot P_{3|G})^\top, \quad (5)$$

$$B = (P_{3|G} \otimes P_{1|H}) \text{diag}(P_{HG}(\cdot)) (P_{4|G} \otimes P_{2|H})^\top, \quad (6)$$

$$C = (P_{4|G} \otimes P_{1|H}) \text{diag}(P_{HG}(\cdot)) (P_{3|G} \otimes P_{2|H})^\top. \quad (7)$$

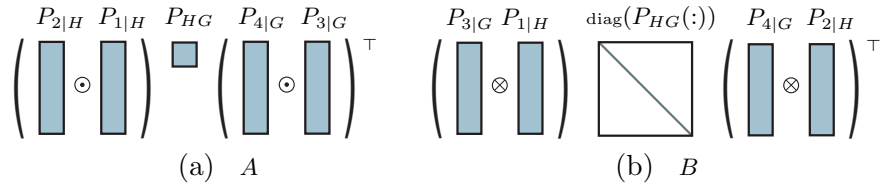


Figure 4: Schematic diagrams of the two unfoldings  $A$  and  $B$ .

The factorization of  $A$  is very different from those of  $B$  and  $C$ . First, in  $A$ ,  $P_{2|H} \odot P_{1|H}$  is a matrix of size  $n^2 \times k$ , and the columns of  $P_{2|H}$  interact only with their corresponding columns in  $P_{1|H}$ . However, in  $B$ ,  $P_{3|G} \otimes P_{1|H}$  is a matrix of size  $n^2 \times k^2$ , and every column of  $P_{1|H}$  interacts with every column of  $P_{3|G}$  respectively (similarly for  $C$ ). Second, in  $A$ , the middle factor  $P_{HG}$  has size  $k \times k$ , whereas in  $B$ , the entire of  $P_{HG}$  appear as the diagonal of a matrix of size  $k^2 \times k^2$  (similarly for  $C$ ). These differences result in different rank properties of  $A$ ,  $B$  and  $C$  which we will exploit to discover the latent structure of a quartet.

### 3.2 Rank Properties of the Unfoldings

Under assumption **(A1)** that all CPTs have full column rank, the factorization of  $A$ ,  $B$  and  $C$  in (5), (6) and (7) respectively suggest that (see appendix §9 for more details)

$$\text{rank}(A) = \text{rank}(P_{HG}) = k \leq \text{rank}(B) = \text{rank}(C) = \text{nnz}(P_{HG}), \quad (8)$$

where  $\text{nnz}(\cdot)$  denotes the number of nonzero elements. We note that the equality is attained if and only if the relationship between the hidden variables  $G$  and  $H$  is deterministic, *i.e.*, there is a single nonzero element in each row and in each column of  $P_{HG}$ . In this case, the grouping of variables in a quartet can be arbitrary, and we will not consider this case in the paper. More specifically, we have

**Theorem 3.** *Assume  $P_{HG}$  has a few zero entries, then  $k \ll k^2 \approx \text{nnz}(P_{HG})$  and thus*

$$\boxed{\text{rank}(A) \ll \text{rank}(B) = \text{rank}(C)}. \quad (9)$$

The above theorem reveals a useful difference between the correct grouping of variables and the two incorrect ones. Furthermore, this condition can be easily verified: Given  $\mathcal{P}$  we can check the rank of its matrix representations  $A$ ,  $B$  and  $C$  and thus discover the latent structure of the quartet.

### 3.3 Nuclear Norm Relaxation for the Rank Condition

In practice, due to sampling noise all unfolding matrices  $A$ ,  $B$  and  $C$  would be nearly full rank, so the rank condition cannot be applied directly. To deal with this, we design a test based on relaxation of the rank condition using nuclear norm

$$\|M\|_* = \sum_{i=1}^n \sigma_i(M), \quad (10)$$

which is the sum of all singular values of an  $(n \times n)$  matrix  $M$ . Instead of comparing the ranks of  $A$ ,  $B$  and  $C$ , we look for the one with the smallest nuclear norm and declare the latent structure corresponding to it. This simple quartet algorithm is summarized in Algorithm 1. Note that

---

**Algorithm 1**  $i^* = \text{Quartet}(X_1, X_2, X_3, X_4)$

---

- 1: Estimate  $\hat{\mathcal{P}}(X_1, X_2, X_3, X_4)$  from a set of  $m$  *i.i.d.* samples  $\{(x_1^l, x_2^l, x_3^l, x_4^l)\}_{l=1}^m$ .
  - 2: Unfold  $\hat{\mathcal{P}}$  in three different ways into matrices  $\hat{A}$ ,  $\hat{B}$  and  $\hat{C}$ , and compute their nuclear norms  $a_1 = \|\hat{A}\|_*$ ,  $a_2 = \|\hat{B}\|_*$  and  $a_3 = \|\hat{C}\|_*$ .
  - 3: Return  $i^* = \text{argmin}_{i \in \{1, 2, 3\}} a_i$ .
- 

Algorithm 1 works even if the number of hidden states,  $k$ , is a priori unknown. This is an important advantage over the idea of learning the structure based on additive distance (Lake, 1994), where  $k$  is assumed to be the same as the number of states,  $n$ , of the observed variables, or over a recent approach based on quartet test (Anandkumar et al., 2011), where  $k$  needs to be specified in advance.

In our current context, nuclear norm has a few useful properties. First, it is the tightest convex lower bound of the rank of a matrix (Fazel et al., 2001). This is why<sup>4</sup> it is meaningful to compare nuclear norms instead of ranks. Second, it is easy to compute: a standard singular value decomposition will do the job. Third, it is robust to estimate. The nuclear norm of a probability matrix  $\hat{A}$  based on samples is nicely concentrated around its population quantity (Rosasco et al., 2010). Given a confidence level  $1 - 2e^{-\tau}$ , an estimate based on  $m$  samples satisfies

$$|\|A\|_* - \|\hat{A}\|_*| = \left| \sum_i \sigma_i(A) - \sum_i \sigma_i(\hat{A}) \right| \leq 2\sqrt{2\tau}/\sqrt{m}. \quad (11)$$

Fourth, the nuclear norm can be viewed as a measure of dependence between two pairs of variables. For instance, if  $A$  corresponds to grouping  $\{\{1, 2\}, \{3, 4\}\}$ ,  $\|A\|_*$  measures the dependence between the compound variables  $\{X_1, X_2\}$  and  $\{X_3, X_4\}$ . In the community of kernel methods,  $A$  is treated as a cross-covariance operator between  $\{X_1, X_2\}$  and  $\{X_3, X_4\}$ , and its spectrum has been used to design various dependence measures, such as Hilbert-Schmidt Independence Criterion, which is the sum of squares of all singular values (Gretton et al., 2005a), and kernel constrained covariance, which only takes the largest singular value (Gretton et al., 2005b). Intuitively, our quartet test

---

<sup>4</sup>Note that  $A$ ,  $B$  and  $C$  consist of the same elements so their Frobenius norms are the same, *i.e.*, the 3 matrices are readily equally “normalized”.

says that: if we group the variables correctly, then cross group dependence should be low, since the groups are separated by two latent variables; however if we group the variables incorrectly, then cross group dependence should be high, since similar variables exist in the two groups.

## 4 Recovery Conditions and Finite Sample Guarantee for Quartets

Since nuclear norm is just a convex lower bound of the rank, there might be situations where the nuclear norm does not satisfy the same relation as the rank. That is, it might happen that  $\text{rank}(A) \leq \text{rank}(B)$  but  $\|A\|_* \geq \|B\|_*$ . In this section, we present sufficient conditions under which nuclear norm returns successful quartet test.

**When latent variables  $H$  and  $G$  are independent**,  $\text{rank}(P_{HG}) = 1$ , since  $P_{HG} = P_H P_G^\top$  ( $P(h, g) = P(h)P(g)$ ). Let  $\{\{1, 2\}, \{3, 4\}\}$  be the correct quartet relation. We can obtain simpler characterizations of the 3 unfoldings of  $\mathcal{P}(X_1, X_2, X_3, X_4)$ , denoted as  $A_\perp$ ,  $B_\perp$  and  $C_\perp$  respectively. Using Lemma 2 and the independence of  $H$  and  $G$ , we have (see appendix, (26)–(27))

$$\begin{aligned} A_\perp &= (P_{2|H} \odot P_{1|H}) P_H P_G^\top (P_{4|G} \odot P_{3|G})^\top \\ &= P_{12}(\cdot) P_{34}(\cdot)^\top, \\ B_\perp &= (P_{3|G} \otimes P_{1|H})(\text{diag}(P_G) \otimes \text{diag}(P_H))(P_{4|G} \otimes P_{2|H})^\top \\ &= P_{34} \otimes P_{12}, \end{aligned} \tag{12}$$

and  $\text{rank}(A_\perp) = 1 \ll \text{rank}(B_\perp)$  which is consistent with Theorem 3. Furthermore, since  $A_\perp$  has only one nonzero singular value, we have  $\|A_\perp\|_* = \|A_\perp\|_F = \|B_\perp\|_F \leq \|B_\perp\|_*$  (using  $\|M\|_F \leq \|M\|_*$  for any matrix  $M$ ). Similarly,  $C_\perp = P_{43} \otimes P_{12}$  and  $\|A_\perp\|_* \leq \|C_\perp\|_*$ . Then we know for sure that the nuclear norm quartet test will return the correct topology.

**When latent variables  $H$  and  $G$  are not independent**, we treat it as perturbation  $\Delta$  away from the independent case, *i.e.*,  $\tilde{P}_{HG} = P_H P_G^\top + \Delta$ . The size of  $\Delta$  quantifies the strength of dependence between  $H$  and  $G$ . Obviously, when  $\Delta$  is small, *e.g.*,  $\Delta = \mathbf{0}$ , we are back to the independence case and it is easy to discover the correct quartet relation; when it is large, *e.g.*,  $\Delta = I - P_H P_G^\top$ ,  $H$  and  $G$  are deterministically related and the different groupings are indistinguishable. The question is how large can  $\Delta$  be while still allowing the nuclear norm quartet test to find the correct latent relation.

First, we require **(A3)**  $\Delta \mathbf{1} = \mathbf{0}$ , and  $\Delta^\top \mathbf{1} = \mathbf{0}$ , where  $\mathbf{1}$  and  $\mathbf{0}$  are vectors of all ones and all zeros. Such perturbation  $\Delta$  keeps the marginal distributions  $P_H$  and  $P_G$  as in the independent case, since  $\tilde{P}_H = \tilde{P}_{HG} \mathbf{1} = P_H P_G^\top \mathbf{1} + \Delta \mathbf{1} = P_H$ . Assuming  $\{\{1, 2\}, \{3, 4\}\}$  is the correct quartet relation,  $\Delta$  also keeps the pairwise marginal distribution  $P_{12}$  as in the independent case, since  $P_{12} = P_{1|H} \text{diag}(P_H) P_{2|H}^\top$  and the marginal  $P_H$  is the same before and after the perturbation. Similar reasoning also applies to  $P_{34} = P_{3|G} \text{diag}(P_G) P_{4|G}^\top$ .

We define *excessive dependence* of the correct and incorrect groupings as

$$\theta := \min\{\|B_\perp\|_* - \|A_\perp\|_*, \|C_\perp\|_* - \|A_\perp\|_*\}.$$

It quantifies the changes in dependence when we switch from incorrect groupings to the correct one (in the case when  $H$  and  $G$  are independent). Note that  $\theta$  is measured only from pairwise marginals (12),  $P_{12}$  and  $P_{34}$ . Using matrix perturbation analysis we can show that (see appendix §11 for proof)

**Lemma 4.** *If  $\|\Delta\|_F \leq \frac{\theta}{k^2+k}$ , then Algorithm 1 returns the correct quartet relation.*

Thus, if the excessive dependence  $\theta$  is large compared to the number of hidden states, the size of the allowable perturbation can be correspondingly larger. In other words, if the dependence between variables within the same group is strong enough compared to the dependence across groups, we allow for larger  $\Delta$  and stronger dependence between hidden variables  $H$  and  $G$  (which is closer to the indistinguishable case). Then under the recovery condition in Lemma 4, and given  $m$  *i.i.d.* observations, we can obtain the following guarantee for the quartet test (see appendix, §13 for proof). Let  $\alpha = \min \{\|B\|_* - \|A\|_*, \|C\|_* - \|A\|_*\}$ .

**Lemma 5.** *With probability  $1 - 8e^{-\frac{1}{32}m\alpha^2}$ , Algorithm 1 returns the correct quartet relation.*

## 5 Building Latent Tree from Quartets

**Algorithm.** We can use the resolved quartet relations (Algorithm 1) to discover the structure of the entire tree via an incremental divide-and-conquer algorithm (Pearl & Tarsi, 1986; Pearl, 1988), summarized in Algorithm 2 (further details in appendix §10). Joining variable  $X_{i+1}$  to the current tree of  $i$  leaves can be done with  $O(\log i)$  tests. This amounts to performing  $O(d \log d)$  quartet tests for building an entire tree of  $d$  leaves, which is efficient even if  $d$  is large. Moreover, as shown in (Pearl & Tarsi, 1986), this algorithm is consistent.

---

**Algorithm 2**  $\mathcal{T} = \text{BuildTree}(X_1, \dots, X_d)$

---

- 1: Connect any 4 variables  $X_1, X_2, X_3, X_4$  with 2 latent variables in a tree  $\mathcal{T}$  using Algorithm 1.
  - 2: **for**  $i = 4, 5, \dots, d - 1$  **do** {insert  $(i+1)$ -th leaf  $X_{i+1}$ }
  - 3:   Choose root  $R$  that splits  $\mathcal{T}$  into sub-trees  $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$  of roughly equal size.
  - 4:   Choose any triplet  $(X_{i_1}, X_{i_2}, X_{i_3})$  of leaves from different sub-trees.
  - 5:   Test which sub-tree should  $X_{i+1}$  be joined to:  
 $i^* \leftarrow \text{Quartet}(X_{i+1}, X_{i_1}, X_{i_2}, X_{i_3})$ .
  - 6:   Repeat recursively from step 3 with  $\mathcal{T} := \mathcal{T}_{i^*}$ .  
This will eventually reduce to a tree with a single leaf. Join  $X_{i+1}$  to it via hidden variable.
  - 7: **end for**
- 

**Tree recovery conditions and guarantees.** How will the quartet recovery conditions translate to recovery conditions for the entire tree, where each “edge” of a quartet is a path in the tree? What are the finite sample guarantees for the divide-and-conquer algorithm?

When a quartet is taken from a latent tree, each edge of the quartet corresponds to a path in the tree involving a chain of variables (Figure 2). We need to bound the perturbation to each single edge of the tree such that joint path perturbations satisfy edge perturbation conditions from Lemma 4. For a quartet  $q = \{\{i_1, i_2\}, \{i_3, i_4\}\}$  corresponding to a single edge between  $H$  and  $G$ , denote the excessive dependence by  $\theta_q$ . By adding perturbation  $\Delta_q$  of size smaller than  $\frac{\theta_q}{k^2+k}$  to  $P_H P_G^\top$  we can still correctly recover  $q$ . Let  $\theta_{\min} := \min_{\text{quartet } q} \theta_q$ . If we require  $\|\Delta_q\|_F \leq \frac{\theta_{\min}}{k^2+k}$ , all such quartet relations will be recovered successfully. If we further restrict the size of the perturbation by the smallest value in a marginal probability distribution of a hidden variable,  $\gamma_{\min} := \min_{\text{hidden node } H} \min_{i=1\dots k} P_H(i)$ , we can guarantee that all quartet relations corresponding



to a path between  $H$  and  $G$  can also be successfully recovered by the nuclear norm test (see appendix §12). Therefore, we assume that **(A4)**  $\|\Delta_q\|_F \leq \min\{\frac{\theta_{\min}}{k^2+k}, \gamma_{\min}\}$  for all quartets  $q$  in a tree.

**Theorem 6.** *Algorithm 2 returns the correct tree topology under assumptions **(A1)**–**(A4)**.*

The recovery conditions guarantee that all quartet relations can be resolved correctly and simultaneously. Then a consistent algorithm using a subset of the quartet relations should return the correct tree structure. Given  $m$  *i.i.d.* samples, we have the following statistical guarantee for the tree building algorithm (see appendix, §14 for proof). Let  $\alpha_{\min} := \min_{\text{quartet } q} \alpha_q$ .

**Theorem 7.** *With probability  $1 - 8 \cdot c \cdot d \log d \cdot e^{-\frac{1}{32} m \alpha_{\min}^2}$ , Algorithm 2 recovers the correct tree topology for a constant  $c$  under assumptions **(A1)**–**(A4)**.*

We note that there are better quartet based algorithms for building latent trees with stronger statistical guarantees, *e.g.* (Erdős et al., 1999). We can adapt our nuclear norm based quartet test to those algorithm as well. However, this is not the main focus of the paper. We choose the divide-and-conquer algorithm due to its simplicity, ease of analysis and it illustrates well how our quartet recovery guarantee can be translated into a tree building guarantee.

## 6 Experiments

We compared our algorithm with representative algorithms: the neighbor-joining algorithm (NJ) (Saitou & Nei, 1987), a quartet based algorithm of Anandkumar et al. (2011) (Spectral@ $k$ ), the Chow-Liu neighbor Joining algorithm (CLNJ) (Choi et al., 2011), and an algorithm of Harmeling & Williams (2010) (HW).

NJ proceeds by recursively joining two variables that are closest according to an additive distance defined as  $d_{ij} = \frac{1}{2} \log \det \text{diag } P_i - \log |\det P_{ij}| + \frac{1}{2} \log \det \text{diag } P_j$ , where “det” denotes determinant, “diag” is a diagonalization operator,  $P_{ij}$  denotes the joint probability table  $P(X_i, X_j)$ , and  $P_i$  and  $P_j$  the probability vector  $P(X_i)$  and  $P(X_j)$  respectively (Lake, 1994). When  $P_{ij}$  has rank  $k < n$ ,  $\log |\det P_{ij}|$  is not defined, NJ can perform poorly. Spectral@ $k$  uses singular values of  $P_{ij}$  to design a quartet test (Anandkumar et al., 2011). For instance, if the true quartet configuration is  $\{\{1, 2\}, \{3, 4\}\}$  as in Figure 2, then the quartet needs to satisfy  $\prod_{s=1}^k \sigma_s(P_{12})\sigma_s(P_{34}) > \max\{\prod_{s=1}^k \sigma_s(P_{13})\sigma_s(P_{24}), \prod_{s=1}^k \sigma_s(P_{14})\sigma_s(P_{23})\}$ . Based on this relation, a confidence interval based quartet test is designed and used as a subroutine for a tree reconstruction algorithm. Spectral@ $k$  can handle cases with  $k < n$ , but still require  $k$  as an input. We will show in later experiments that its performance is sensitive to the choice of  $k$ . CLNJ first applies Chow-Liu algorithm (Chow & Liu, 1968) to obtain a fully observed tree and then proceeds by adding latent variables using neighbor joining algorithm. The HW algorithm is a greedy algorithm to learn binary trees by iteratively joining two nodes with a high mutual information. The number of hidden states is automatically determined in the HW algorithm and can be different for different latent variables.

### 6.1 Resolving Quartet Relations

We compared our method to NJ and Spectral@ $k$  in terms of their ability to recover the quartet relation among four variables. We used quartet with three different configurations for the hidden states: (1)  $k_H = 2$  and  $k_G = 4$  (small difference); (2)  $k_H = 2$ ,  $k_G = 8$  (large difference); and (3)

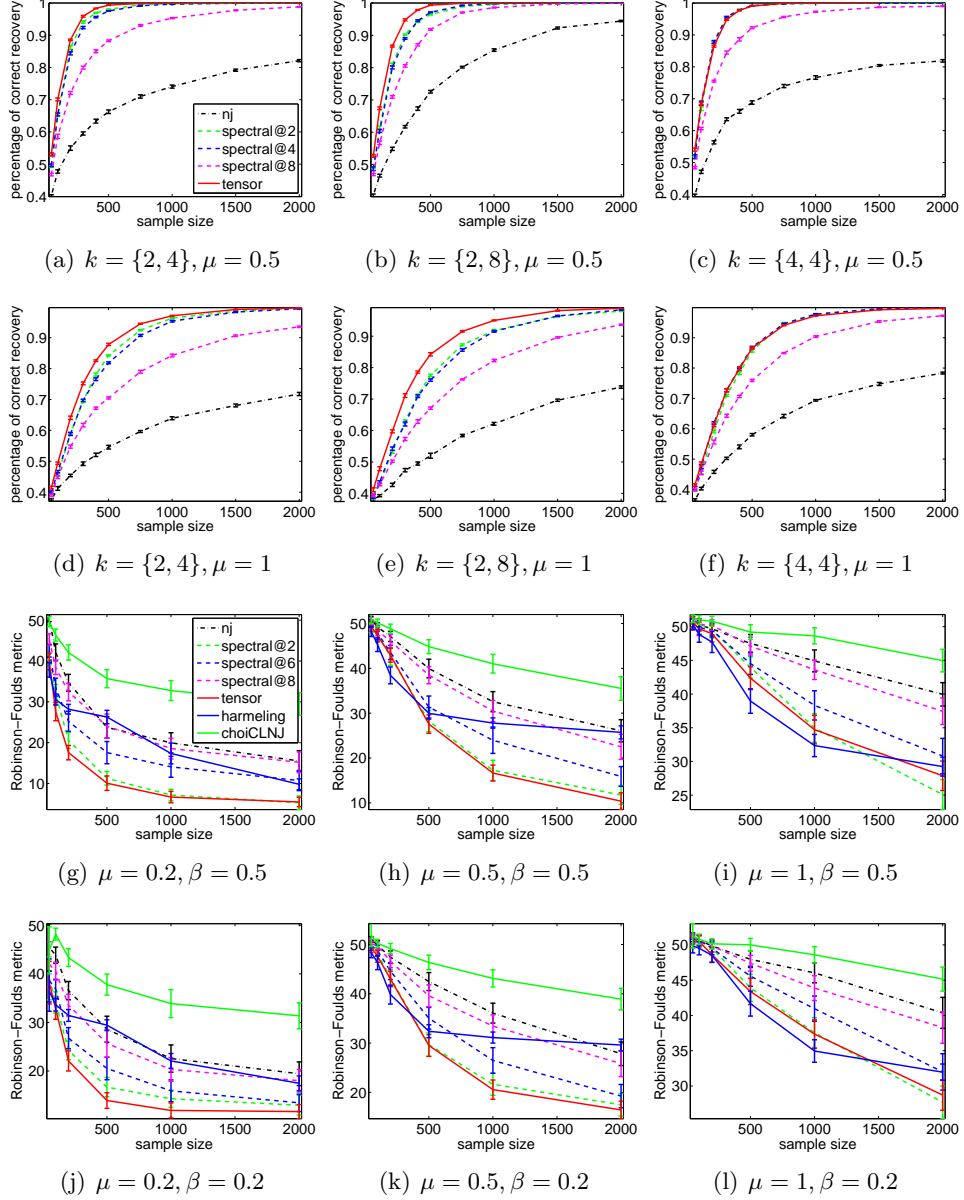


Figure 5: (a)-(f) Quartet recovery results. (g)-(l) Tree recovery results. “tensor” is our method.

$k_H = 4, k_G = 4$  (no difference). In all cases, the states of the observed variables were fixed to  $n = 10$ . In all cases we started from independent  $P_{HG}$  but identity  $P_{X_i|H}$  and  $P_{X_i|G}$ , and perturbed them using the following formula  $P(a = i|b) = \frac{P(a=i|b)+u_i}{\sum_i P(a=i|b)+u_i}$ , where all  $u_i$  are *i.i.d.* random variables drawn from  $\text{Uniform}[0, \mu]$ . We then drew random sample from the quartet according to these CPTs. We studied the percentage of correctly recovered quartet relations as we varied the sample size across  $S = \{50, 100, 200, 300, 400, 500, 750, 1000, 1500, 2000\}$  and under two different levels of perturbation ( $\mu = \{0.5, 1\}$ ). We randomly initialized each experiment 1000 times and report the average quartet recovery performance and the standard error in Figure 5.

The proposed method compares favorably to NJ and Spectral@ $k$ . The performance of Spectral@ $k$

varies a lot depending on the chosen number of singular values  $k$ . Our method is free from tuning parameters and often stays among the top performing ones. Especially when the number of hidden states are very different from each other ( $k_H = 2$  and  $k_G = 8$ ), our method is leading the second best by a large gap (Figure 5(b) and 5(e)). When both hidden states are the same ( $k_H = k_G = 4$ ), the Spectral@ $k$  achieves the best performance when the chosen number of singular values  $k$  is the same as  $k_H$ . Note that allowing Spectral@ $k$  to use different  $k$  resembles using cross validations for finding the best  $k$ . It is expensive while our approach performs almost indistinguishable from Spectral@ $k$  even it choose the best  $k$ .

## 6.2 Discovering Latent Tree Structure

We used different tree topologies and sample sizes in this experiment. We generated tree topologies by randomly splitting 16 observed variables recursively into two groups. The recursive splitting stops when there are only two nodes left in a group. We introduced a hidden variable to join the two partitions in each recursion and this gives a latent tree structure. The topology of the tree is controlled by a single splitting parameter  $\beta$  which controls the relative size of the first partition versus the second. If  $\beta$  is close to 0 or 1, we obtain trees of skewed shape, with long path of hidden variables. If  $\beta$  is close to 0.5, the resulting latent trees are more balanced. In our experiments, we experimented with skewed latent trees  $\beta = 0.2$  and balanced trees  $\beta = 0.5$ . We first generate different random  $k$  between 2 and 8 for the hidden states, and then generate the probability models for each tree using the same scheme as in our previous experiment. Here we experimented with perturbation level  $\mu = \{0.2, 0.5, 1\}$ .

We varied the sample size across  $S = \{50, 100, 200, 500, 1000, 2000\}$ , and measured the error of the constructed tree using Robinson-Foulds metric (Robinson & Foulds, 1981). This measure is a metric over trees of the same number of leaves. It is defined as  $(a + b)$  where  $a$  is the number of partitions of variables implied by the learned tree but not by the true tree and  $b$  is the number of partitions of the variables implied by the true tree but not by the learned tree (in a sense similar to precision and recall score).

The tree recovery results are shown in Figure 5(g)-5(l). Again we can see that our proposed method compares favorably to existing algorithms. All through the 6 experimental conditions, the tensor approach and spectral@2 performed the best with sufficiently large sample sizes. Note that we tried out different  $k$  for Spectral@ $k$  which resembles using cross validations for finding the best  $k$ . Even in this case, our approach works comparably without having to know  $k$ . Harmeling-William’s algorithm performed well in small sample sizes, while CLNJ does not perform well in these experimental conditions.

## 6.3 Understanding Latent Relations between Stocks

We applied our algorithm to discover a latent tree structure from a stock dataset. Our goal is to understand how stock prices  $X_i$  are related to each other. We acquired closing prices of 59 stocks from 1984 to 2011 (from [www.finance.yahoo.com](http://www.finance.yahoo.com)), which provides us 6800 samples. The daily change of each stock price is discretized into 10 values, and we applied our algorithm to build a latent tree. A visualization of the learned tree topologies and discovered groupings are shown in Figure 6.

We see nice groupings of stocks according to their industrial sectors. For instance, companies related to petroleum, such as CVX (Chevron), XOM (Exxon Mobil), APA (Apache), COP (Cono-

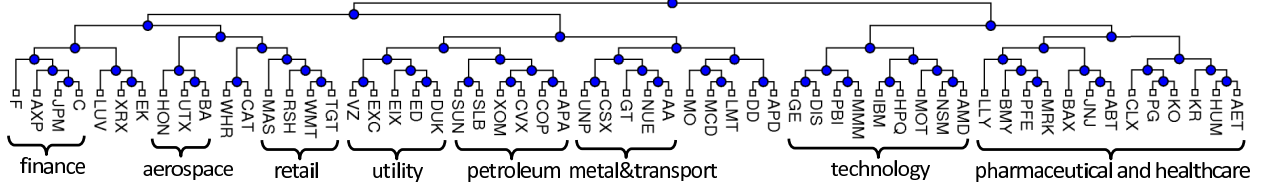


Figure 6: Latent tree estimated from stock data.

coPhillips), SLB (Schlumberger) and SUN (Sunoco), are grouped into a subtree. Pharmaceutical companies, such as MRK (Merck), PFE (Pfizer), BMY (Bristol Myers Squibb), LLY (Eli Lilly), ABT (Abbott Laboratories), JNJ (Johnson and Johnson) and BAX (Baxter International), are all grouped into a subtree. High-tech companies, such as AMD, MOT (Motorola), HPQ (Hewlett-Packard), IBM, are grouped into another subtree. There are also subtree for retailers, such as TGT (Target), WMT (Wal-Mart), RSH (RadioShack), subtree for utility service companies, such as DUK (Duke Energy), ED (Consolidated Edison), EIX (Edison), ECX (Exelon), VZ (Verizon), and subtree related to financial companies, such as C (Citigroup), JPM (JPMorgan Chase), and AXP (American Express). We can also see subtree related to financial companies, such as C (Citigroup), JPM (JPMorgan Chase), and AXP (American Express). An interesting observation is that F (Ford Motor) which is well-known for its car manufacturing is also placed in the same branch as these financial companies. This seemingly abnormal structure can be explained by the fact that Ford Motor operates under two segments: Automotive and Financial Services. Its financial services include the operations of Ford Motor Credit Company and other financial services including holding companies, and real estate. In this respect, it is quite interesting that our algorithm discovered this hidden information.

We also compared different algorithms in terms of held-out likelihood. We first randomized the data 10 times, and each time used half for training and half for computing the held-out likelihood. Then we estimated the latent binary tree structures using different algorithms. Finally, we fit latent variable models to the discovered structures. The number of the states for all hidden variables,  $k$ , were the same in each latent variable model. We experimented with  $k = 2, 4, 6, 8, 10$  to simulate the process of using cross validation to select the best  $k$ . The results are presented in Table 1. Note

Table 1: Negative log-likelihood ( $\times 10^5$ ) on test data. The small the number the better the method.

	Tensor	Spectral@ $k$	Choi (CLNJ)	Neighbor-joining	Harmeling	Chow-Liu
$k = 2$	4.41	4.44	4.43	4.43	4.31	4.41
$k = 4$	4.30	4.35	4.33	4.33		
$k = 6$	4.28	4.35	4.32	4.31		
$k = 8$	<b>4.28</b>	4.35	4.32	4.31		
$k = 10$	4.29	4.37	4.32	4.31		

that Harmeling-William’s algorithm automatically discovers  $k$ , so it does not use the experimental parameter  $k$ . Chow-Liu tree does not contain any hidden variables and hence just one number in the table. CLNJ and Neighbor-joining assume the states for the hidden and observed variables are the same during structure learning. However, in parameter fitting, we can still use different number

of hidden states  $k$ . In this experiment, the structure produced by our tensor approach produced the best held-out likelihood.

## 7 Conclusion

In this paper, we propose a quartet-based method for discovering the tree structures of latent variable models. The practical advantage of the new method is that we do not need to pre-specify the number of the hidden states, a quantity usually unknown in practice. The key idea is to view the joint probability tables of quadruple of variables as 4th order tensors and then use the spectral properties of the unfolded tensors to design a quartet test. We provide conditions under which the algorithm is consistent and its error probability decays exponentially with increasing the sample size. In both simulated and a real dataset, we demonstrated the usefulness of our methods for discovering latent structures. While in this study we focus on the properties of the 4th order tensor and its various unfoldings, we believe that properties of tensors and methods and algorithms from multilinear algebra will allow to address many other problems arising from latent variable models.

## References

- Allman, E. S. and Rhodes, J. A. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *Journal of Computational Biology*, 13(5):1101–1113, 2006.
- Anandkumar, A., Chaudhuri, K., Hsu, D., Kakade, S., Song, L., and Zhang, T. Spectral methods for learning multivariate latent tree structure. In *Neural Information Processing Systems*, 2011.
- Buneman, P. The recovery of trees from measures of dissimilarity. In Hodson, F.R., Kendall, D.G., and Tautu, P. (eds.), *Mathematics in the archaeological and historical sciences*, pp. 387–395. Edinburgh University Press, 1971.
- Carroll, J. and Chang, J. Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. *Psychometrika*, 35(3):283–319, 1970.
- Choi, M., Tan, V., Anandkumar, A., and Willsky, A. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.
- Chow, C., and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.
- Erdős, P. L., Székely, L. A., Steel, M. A., and Warnow., T. J. A few logs suffice to build (almost) all trees: Part II. *Theoretical Computer Science*, 221:77–118, 1999.
- Eriksson, N. Tree construction using singular value decomposition. In Pachter, L. and Sturmfels, B. (eds.), *Algebraic Statistics for Computational Biology*, pp. 347–358. Cambridge University Press, 2005. URL <http://dx.doi.org/10.1017/CB09780511610684>.
- Fazel, Maryam, Hindi, Haitham, and Boyd, Stephen P. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, pp. 4734–4739, 2001.

- Grasedyck, L. Hierarchical singular value decomposition of tensors. *SIAM J. Matrix Anal. Appl.*, 31(4):2029–2054, 2010.
- Gretton, A., Bousquet, O., Smola, A. J., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In Jain, S., Simon, H. U., and Tomita, E. (eds.), *Proceedings of the International Conference on Algorithmic Learning Theory*, pp. 63–77. Springer-Verlag, 2005a.
- Gretton, A., Herbrich, R., Smola, A. J., Bousquet, O., and Schölkopf, B. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b.
- Harmeling, S. and Williams, C. Greedy learning of binary latent trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1087–1097, 2010.
- Harshman, R. A. Foundations of the PARAFAC procedure: Model and conditions for an “explanatory” multi-mode factor analysis. *UCLA Working Papers in Phonetics*, 16(1):1–84, 1970.
- Heller, K. A. and Ghahramani, Z. Bayesian hierarchical clustering. In *Proceedings of the International Conference on Machine Learning*, pp. 297–304, 2005.
- Lake, J.A. Reconstructing evolutionary trees from dna and protein sequences: paralineal distances. *Proceedings of the National Academy of Sciences*, 91(4):1455, 1994.
- Mihaescu, R., Levy, D., and Pachter, L. Why neighbor-joining works. *Algorithmica*, 54(1):1–24, 2009.
- Oseledets, I. V. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33:2295–2317, 2011.
- Parikh, A., Song, L., and Xing, E. P. A spectral algorithm for latent tree graphical models. In *Proceedings of the International Conference on Machine Learning*, 2011.
- Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, 1988.
- Pearl, J. and Tarsi, M. Structuring causal trees. *Journal of Complexity*, 2(1):60–77, 1986.
- Robinson, D.F. and Foulds, L.R. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.
- Rosasco, L., Belkin, M., and Vito, E.D. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.
- Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.
- Semple, C. and Steel, M.A. *Phylogenetics*, volume 24. Oxford University Press, USA, 2003.
- Teh, Yee Whye, Daume, Hal, and Roy, Daniel. Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems 22*, 2008.
- Zhang, N. L. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723, 2004.

# Unfolding Latent Tree Structures using 4th Order Tensors

## Appendix

### 8 Properties and Notations used

Nuclear and Frobenius norms:

- Let  $\sigma_i$  be the singular values of  $A$ . Then

$$\|A\|_* = \sum_i \sigma_i, \quad \|A\|_F^2 = \sum_i \sigma_i^2 \quad \text{and} \quad \|A\|_F \leq \|A\|_*. \quad (13)$$

- (Nuclear and Frobenius norms are unitarily invariant) For any orthogonal  $Q$  we have

$$\begin{aligned} \|A\|_* &= \|QA\|_* = \|AQ\|_*, \\ \|A\|_F &= \|QA\|_F = \|AQ\|_F. \end{aligned} \quad (14)$$

- $\|AB\|_* \leq \|A\|_F \|B\|_F \leq \|A\|_* \|B\|_*$ .
- Let  $\sigma_i$  be the singular values of  $X$  and  $\tilde{\sigma}_i$  be the singular values of  $\tilde{X} = X + E$ . Then

$$\|\text{diag}(\tilde{\sigma}_i - \sigma_i)\|_* \leq \|\tilde{X} - X\|_*. \quad (15)$$

Kronecker and Khatri-Rao products:

$$(A \otimes B)^\top = A^\top \otimes B^\top \quad (16)$$

$$(A + B) \otimes C = A \otimes C + B \otimes C \quad (17)$$

$$AB \otimes CD = (A \otimes C)(B \otimes D) \quad (18)$$

$$AB \odot CD = (A \otimes C)(B \odot D) \quad (19)$$

$$\|A \otimes B\|_F = \|A\|_F \|B\|_F$$

$$\text{rank}(A \otimes B) = \text{rank}(A) \text{rank}(B)$$

Tensor operations:

We use the following tensor-matrix products of a tensor  $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$  with matrices  $M^{(n)} \in \mathbb{R}^{J_n \times I_n}$ ,  $n = 1, 2, 3$ :

$$\text{mode-1 product: } (\mathcal{A} \bullet_1 M^{(1)})_{j_1 i_2 i_3} = \sum_{i_1=1}^{I_1} a_{i_1 i_2 i_3} m_{j_1 i_1}^{(1)},$$

$$\text{mode-2 product: } (\mathcal{A} \bullet_2 M^{(2)})_{i_1 j_2 i_3} = \sum_{i_2=1}^{I_2} a_{i_1 i_2 i_3} m_{j_2 i_2}^{(2)},$$

$$\text{mode-3 product: } (\mathcal{A} \bullet_3 M^{(3)})_{i_1 i_2 j_3} = \sum_{i_3=1}^{I_3} a_{i_1 i_2 i_3} m_{j_3 i_3}^{(3)},$$

where  $1 \leq i_n \leq I_n$ ,  $1 \leq j_n \leq J_n$ . These products can be considered as a generalization of the left and right multiplication of a matrix  $A$  with a matrix  $M$ . The mode-1 product signifies multiplying

the columns (mode-1 vectors) of  $\mathcal{A}$  with the rows of  $M^{(1)}$  and similarly for the other tensor-matrix products.

The *contracted product*  $\mathcal{C}$  of two tensors  $\mathcal{A} \in \mathbb{R}^{I \times J \times M}$  and  $\mathcal{B} \in \mathbb{R}^{K \times L \times M}$  along their third modes is a 4th order tensor denoted by  $\mathcal{C} = \langle \mathcal{A}, \mathcal{B} \rangle_3$ .  $\mathcal{C} \in \mathbb{R}^{I \times J \times K \times L}$  and its entries  $\mathcal{C}(i, j, k, l)$ ,  $1 \leq i \leq I$ ;  $1 \leq j \leq J$ ;  $1 \leq k \leq K$ ;  $1 \leq l \leq L$  are defined as

$$\mathcal{C}(i, j, k, l) = \sum_{m=1}^M a_{ijm} b_{klm}.$$

It can be interpreted as taking inner products of the mode-3 vectors of  $\mathcal{A}$  and  $\mathcal{B}$  and storing the results in  $\mathcal{C}$ .

The 3 different *reshapings*  $A$ ,  $B$  and  $C$  (2)–(4) of the tensor  $\mathcal{P}$  contain exactly the same entries as  $\mathcal{P}$  but in different order.

- $A$  corresponds to the grouping  $\{\{1, 2\}, \{3, 4\}\}$  of the variables. The rows of  $A$  correspond to dimensions 1 and 2 of  $\mathcal{P}$ , and its columns to dimensions 3 and 4. Suppose all observed variables take values from  $\{1, \dots, n\}$ , then entry of  $A$  at  $(x_1 + n(x_2 - 1))$ -th row and  $(x_3 + n(x_4 - 1))$ -th column is equal to  $\mathcal{P}(x_1, x_2, x_3, x_4)$ ;
- $B$  corresponds to the grouping  $\{\{1, 3\}, \{2, 4\}\}$ , and its entry at  $(x_1 + n(x_3 - 1))$ -th row and  $(x_2 + n(x_4 - 1))$ -th column is equal to  $\mathcal{P}(x_1, x_2, x_3, x_4)$ ;
- $C$  corresponds to the grouping  $\{\{1, 4\}, \{2, 3\}\}$ , and its entry at  $(x_1 + n(x_4 - 1))$ -th row and  $(x_2 + n(x_3 - 1))$ -th column is equal to  $\mathcal{P}(x_1, x_2, x_3, x_4)$ .

## 9 Matrix Representations $A$ , $B$ , $C$ of $\mathcal{P}$

From  $\mathcal{P}$  to  $A$ ,  $B$ ,  $C$ :

Let  $X \in \mathbb{R}^{m \times k}$ ,  $Y \in \mathbb{R}^{k \times l}$ ,  $Z \in \mathbb{R}^{n \times l}$ ,  $X = (x_1, \dots, x_k)$  and  $Z = (z_1, \dots, z_l)$ . A useful property that we will use in our derivations is the following

$$X Y Z^\top = \sum_{i,j} x_i y_{ij} z_j^\top. \quad (20)$$

We can derive the formula for  $A$  starting from the element-wise formula (1)

$$\mathcal{P}(x_1, x_2, x_3, x_4) = \sum_{h,g} P(x_1|h)P(x_2|h)P(h,g)P(x_3|g)P(x_4|g)$$

and placing all entries in the matrix  $A$  in the correct order. Note that given  $h$  and  $g$  we only need one column of each  $P_{1|H}$ ,  $P_{2|H}$ ,  $P_{3|G}$  and  $P_{4|G}$ , which we will denote by  $(P_{1|H})_h$ ,  $(P_{2|H})_h$ ,  $(P_{3|G})_g$  and  $(P_{4|G})_g$ . In order to obtain a matrix such that  $X_1$  and  $X_2$  are mapped to rows and  $X_3$  and  $X_4$  are mapped to columns, we need to map all possible products of single element of  $(P_{1|H})_h$  and single element of  $(P_{2|H})_h$  to rows and similarly, we need to map all possible products of single element of  $(P_{3|G})_g$  and single element of  $(P_{4|G})_g$  to columns. This can be done using Khatri-Rao products in the following way

$$\begin{aligned} A &= \sum_{h,g} \left( (P_{2|H})_h \odot (P_{1|H})_h \right) (P_{HG})_{hg} \left( (P_{4|G})_g \odot (P_{3|G})_g \right)^\top \\ &\stackrel{(20)}{=} (P_{2|H} \odot P_{1|H}) P_{HG} (P_{4|G} \odot P_{3|G})^\top. \end{aligned}$$



The matrix  $B$  is unfolding of  $\mathcal{P}$ , such that the rows of  $B$  correspond to  $X_1$  and  $X_3$  and the columns of  $B$  correspond to  $X_2$  and  $X_4$ . We have

$$\begin{aligned}
B &= \sum_{h,g} \left( (P_{3|G})_g \odot (P_{1|H})_h \right) (P_{HG})_{hg} \left( (P_{4|G})_g \odot (P_{2|H})_h \right)^\top \\
&\stackrel{(16)}{=} \sum_{h,g} \left( (P_{3|G})_g \otimes (P_{1|H})_h \right) (P_{HG})_{hg} \left( (P_{4|G})_g^\top \otimes (P_{2|H})_h^\top \right) \\
&\stackrel{(18)}{=} \sum_{h,g} (P_{HG})_{hg} \left( (P_{3|G})_g (P_{4|G})_g^\top \right) \otimes \left( (P_{1|H})_h (P_{2|H})_h^\top \right) \\
&\stackrel{(17)}{=} \sum_h \left( \sum_g (P_{HG})_{hg} (P_{3|G})_g (P_{4|G})_g^\top \right) \otimes \left( (P_{1|H})_h (P_{2|H})_h^\top \right) \\
&\stackrel{(20)}{=} \sum_h \left( P_{3|G} \text{diag}((P_{HG})_h) P_{4|G}^\top \right) \otimes \left( (P_{1|H})_h (P_{2|H})_h^\top \right) \\
&\stackrel{(18)}{=} \sum_h \left( P_{3|G} \otimes (P_{1|H})_h \right) \text{diag}((P_{HG})_h) \left( P_{4|G}^\top \otimes (P_{2|H})_h^\top \right) \\
&\stackrel{\text{block-(20)}}{=} \left( P_{3|G} \otimes P_{1|H} \right) \text{diag}(P_{HG}(:)) \left( P_{4|G}^\top \otimes P_{2|H}^\top \right) \\
&\stackrel{(16)}{=} \left( P_{3|G} \otimes P_{1|H} \right) \text{diag}(P_{HG}(:)) \left( P_{4|G} \otimes P_{2|H} \right)^\top.
\end{aligned}$$

The expression for  $C$  is derived in a similar way.

Other representations of  $A$ ,  $B$ ,  $C$ :

Using the properties in Section 8 and the formulas (5)–(7) for the matrix unfoldings  $A$ ,  $B$  and  $C$ , we can derive the following additional formulas,

$$\begin{aligned}
A &= (P_{2|H} \odot P_{1|H}) P_{HG} (P_{4|G} \odot P_{3|G})^\top \\
&= (I_n P_{2|H} \odot P_{1|H} I_H) P_{HG} (I_n P_{4|G} \odot P_{3|G} I_G)^\top \\
&\stackrel{(19)}{=} (I_n \otimes P_{1|H}) (P_{2|H} \odot I_H) P_{HG} (P_{4|G} \odot I_G)^\top (I_n \otimes P_{3|G})^\top \\
&= \begin{pmatrix} P_{1|H} & & & \\ & \ddots & & \\ & & P_{1|H} & \end{pmatrix} \begin{pmatrix} p_{2|H}^{(1,1)} & & & \\ & p_{2|H}^{(1,2)} & & \\ & & \ddots & \\ p_{2|H}^{(2,1)} & & & \\ \vdots & \ddots & & \end{pmatrix} P_{HG} \begin{pmatrix} p_{4|G}^{(1,1)} & & & \\ & p_{4|G}^{(1,2)} & & \\ & & \ddots & \\ p_{4|G}^{(2,1)} & & & \\ \vdots & \ddots & & \end{pmatrix}^\top \begin{pmatrix} P_{3|G} & & & \\ & \ddots & & \\ & & P_{3|G} & \end{pmatrix}^\top, \tag{21}
\end{aligned}$$

$$\begin{aligned}
B &= (P_{3|G} \otimes P_{1|H}) \text{diag}(P_{HG}(:)) (P_{4|G} \otimes P_{2|H})^\top \\
&= (P_{3|G} I_G \otimes I_n P_{1|H}) \text{diag}(P_{HG}(:)) (P_{4|G} I_G \otimes I_n P_{2|H})^\top \\
&\stackrel{(18),(16)}{=} \begin{pmatrix} P_{3|G} \otimes I_n \\ I_G \otimes P_{1|H} \end{pmatrix} \text{diag}(P_{HG}(:)) \begin{pmatrix} I_G \otimes P_{2|H} \\ P_{4|G} \otimes I_n \end{pmatrix}^\top \\
&= \begin{pmatrix} (p_{3|G}^{(1,1)}) & \dots \\ (p_{3|G}^{(2,1)}) & \\ \vdots & \end{pmatrix} \begin{pmatrix} P_{1|H} & \\ & \ddots \\ & P_{1|H} \end{pmatrix} \text{diag}(P_{HG}(:)) \begin{pmatrix} P_{2|H} & \\ & \ddots \\ & P_{2|H} \end{pmatrix}^\top \begin{pmatrix} (p_{4|G}^{(1,1)}) & \dots \\ (p_{4|G}^{(2,1)}) & \\ \vdots & \end{pmatrix}^\top, \tag{22}
\end{aligned}$$

where  $(p^{(i,j)})$  is a diagonal block of size  $(n \times n)$  with all diagonal elements equal to  $p^{(i,j)}$ .

The formula for  $C$  can be obtained from the ones for  $B$  by swapping the positions of  $P_{3|G}$  and  $P_{4|G}$ .

### Rank properties of $A, B, C$ :

In this section we prove the rank properties used in Section 3.2 of the paper.

**Lemma.** *If  $X \in \mathbb{R}^{m \times n}$ ,  $Y \in \mathbb{R}^{n \times k}$ ,  $Z \in \mathbb{R}^{l \times m}$ ,  $Y$  has full row rank, and  $Z$  has full column rank, then*

$$\text{rank}(XY) = \text{rank}(X),$$

$$\text{rank}(ZX) = \text{rank}(X).$$

We assume that all CPTs have full column (or row) rank. Then the first two matrices in (21) also have full column rank. The last two matrices have full row rank. From the lemma, it follows that

$$\text{rank}(A) = \text{rank}(P_{HG}) = k \tag{23}$$

Analogously, the first two matrices in (22) have full column rank. The last two matrices have full row rank. From the lemma, it follows that

$$\text{rank}(B) = \text{nnz}(P_{HG}), \tag{24}$$

i.e., generically,

$$\text{rank}(B) = k^2.$$

## 10 Algorithms

---

**Algorithm 3**  $\mathcal{T}_{next} = \text{QuartetTree}(\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, X_4)$

---

**Require:**  $\text{Leaf}(\mathcal{T})$ : leaves of a tree  $\mathcal{T}$ ;

- 1: **for**  $j = 1$  to 3 **do**
  - 2:    $X_i \leftarrow$  Randomly choose a variable from  $\text{Leaf}(\mathcal{T}_i)$
  - 3: **end for**
  - 4:  $i^* \leftarrow \text{Quartet}(X_1, X_2, X_3, X_4), \quad \mathcal{T}_{next} \leftarrow \mathcal{T}_{i^*}$
-

---

**Algorithm 4**  $\mathcal{T} = \text{Insert}(\mathcal{T}, \tilde{\mathcal{T}}, X_i)$ 

---

**Require:**  $\text{Left}(\mathcal{T})$  and  $\text{Right}(\mathcal{T})$ : left and right child branch of the root respectively;  $\mathcal{T} + \mathcal{T}'$ : return a new tree connecting the root of two trees by an edge and use the root of  $\mathcal{T}$  as the new root

- 1: **if**  $|\text{Leaf}(\mathcal{T})| = 1$  **then**
- 2:    $\mathcal{T} \leftarrow$  Form a tree with root  $R$  connecting  $\text{Leaf}(\mathcal{T})$  and  $X_i$ .
- 3: **else**
- 4:    $\mathcal{T}_{next} \leftarrow \text{QuartetTree}(\text{Left}(\mathcal{T}), \text{Right}(\mathcal{T}), \tilde{\mathcal{T}}, X_i)$
- 5:   **if**  $\mathcal{T}_{next} = \text{Left}(\mathcal{T})$  **then**
- 6:      $\mathcal{T} \leftarrow \text{Insert}(\mathcal{T}_{next}, \text{Right}(\mathcal{T}) + \tilde{\mathcal{T}}, X_i)$
- 7:   **else if**  $\mathcal{T}_{next} = \text{Right}(\mathcal{T})$  **then**
- 8:      $\mathcal{T} \leftarrow \text{Insert}(\mathcal{T}_{next}, \text{Left}(\mathcal{T}) + \tilde{\mathcal{T}}, X_i)$
- 9:   **end if**
- 10: **end if**
- 11:  $\mathcal{T} \leftarrow \mathcal{T} + \tilde{\mathcal{T}}$

---

---

**Algorithm 5**  $\mathcal{T} = \text{BuildTree}(\{X_1, \dots, X_d\})$ 

---

- 1: Randomly choose  $X_1, X_2, X_3$  and  $X_4$
- 2:  $i^* \leftarrow \text{Quartet}(X_1, X_2, X_3, X_4)$
- 3:  $\mathcal{T} \leftarrow$  Form a tree with two connecting hidden variables  $H$  and  $G$ , where  $H$  joins  $X_{i^*}$  and  $X_4$ , while  $G$  joins variables in  $\{X_1, X_2, X_3\} \setminus \{X_{i^*}\}$
- 4: **for**  $i = 5$  to  $d$  **do**
- 5:   Pick a root  $R$  from  $\mathcal{T}$  which split it to three branches of equal sizes, and  $\mathcal{T}_{next} \leftarrow \text{QuartetTree}(\text{Left}(\mathcal{T}), \text{Right}(\mathcal{T}), \text{Middle}(\mathcal{T}), X_i)$
- 6:   **if**  $\mathcal{T}_{next} = \text{Left}(\mathcal{T})$  **then**
- 7:      $\mathcal{T} \leftarrow \text{Insert}(\mathcal{T}_{next}, \text{Right}(\mathcal{T}) + \text{Middle}(\mathcal{T}), X_i)$
- 8:   **else if**  $\mathcal{T}_{next} = \text{Right}(\mathcal{T})$  **then**
- 9:      $\mathcal{T} \leftarrow \text{Insert}(\mathcal{T}_{next}, \text{Left}(\mathcal{T}) + \text{Middle}(\mathcal{T}), X_i)$
- 10:   **else if**  $\mathcal{T}_{next} = \text{Middle}(\mathcal{T})$  **then**
- 11:      $\mathcal{T} \leftarrow \text{Insert}(\mathcal{T}_{next}, \text{Right}(\mathcal{T}) + \text{Left}(\mathcal{T}), X_i)$
- 12:   **end if**
- 13: **end for**

---

## 11 Recovery Conditions for Quartet

**Latent variables  $H$  and  $G$  are independent.** In this case,  $\text{rank}(P_{HG}) = 1$ , since  $P(h, g) = P(h)P(g)$ . Applying the relation in Equation 8, we have that  $\text{rank}(A) = 1 \ll \text{rank}(B)$ . Furthermore, since  $A$  has only one nonzero singular value, we have  $\|A\|_* = \|A\|_F = \|B\|_F \leq \|B\|_*$ , since  $\|M\|_F \leq \|M\|_*$  for any  $M$ . In this case, we know for sure that the nuclear norm quartet test will return the correct topology.

**Latent variables  $H$  and  $G$  are not independent.** We analyze this case by treating it as perturbation  $\Delta$  away from the  $P_{HG}$  in the independent case. We want to characterize how large  $\Delta$  can be while still allowing the nuclear norm quartet test to find the correct latent relation. Suppose  $A_\perp$  and  $B_\perp$  are the unfolding matrices in the case where  $H$  and  $G$  are independent. Suppose we add perturbation  $\Delta$  to  $P_{HG}$ , then  $A_\perp = (P_{2|H} \odot P_{1|H}) P_{HG} (P_{4|G} \odot P_{3|G})^\top$  and its perturbed version is

$A = (P_{2|H} \odot P_{1|H}) (P_{HG} + \Delta) (P_{4|G} \odot P_{3|G})^\top$ . We want to bound the difference  $|\|A_\perp\|_* - \|A\|_*|$ . We have

$$\begin{aligned}
|\|A_\perp\|_* - \|A\|_*| &= \left| \sum_i \sigma_i(A_\perp) - \sum_i \sigma_i(A) \right| \\
&\leq \sum_i |\sigma_i(A_\perp) - \sigma_i(A)| \\
&\stackrel{(15)}{\leq} \|A_\perp - A\|_* \\
&\leq \|(P_{2|H} \odot P_{1|H}) \Delta (P_{4|G} \odot P_{3|G})^\top\|_* \\
&\leq \|P_{2|H} \odot P_{1|H}\|_F \|\Delta\|_F \|P_{4|G} \odot P_{3|G}\|_F \\
&\leq k \|\Delta\|_F,
\end{aligned}$$

since  $P_{2|H} \odot P_{1|H}$  and  $P_{4|G} \odot P_{3|G}$  are CPTs with  $k$  columns each, and thus  $\|P_{2|H} \odot P_{1|H}\|_F^2 \leq k$  and  $\|P_{4|G} \odot P_{3|G}\|_F^2 \leq k$ .

Analogously,  $B_\perp = (P_{3|G} \otimes P_{1|H}) \text{diag}(P_{HG}(\cdot)) (P_{4|G} \otimes P_{2|H})^\top$  and its perturbed version is  $B = (P_{3|G} \otimes P_{1|H}) \text{diag}(P_{HG}(\cdot) + \Delta(\cdot)) (P_{4|G} \otimes P_{2|H})^\top$ . We want to bound the difference  $|\|B_\perp\|_* - \|B\|_*|$ . We have

$$\begin{aligned}
|\|B_\perp\|_* - \|B\|_*| &= \left| \sum_i \sigma_i(B_\perp) - \sum_i \sigma_i(B) \right| \\
&\leq \sum_i |\sigma_i(B_\perp) - \sigma_i(B)| \\
&\stackrel{(15)}{\leq} \|B_\perp - B\|_* \\
&\leq \|(P_{3|G} \otimes P_{1|H}) \text{diag}(\Delta(\cdot)) (P_{4|G} \otimes P_{2|H})^\top\|_* \\
&\leq \|P_{3|G} \otimes P_{1|H}\|_F \|\text{diag}(\Delta(\cdot))\|_F \|P_{4|G} \otimes P_{2|H}\|_F \\
&\leq k^2 \|\text{diag}(\Delta(\cdot))\|_F \\
&= k^2 \|\Delta\|_F,
\end{aligned}$$

since  $P_{3|G} \otimes P_{1|H}$  and  $P_{4|G} \otimes P_{2|H}$  are CPTs with  $k^2$  columns, and thus  $\|P_{3|G} \otimes P_{1|H}\|_F^2 \leq k^2$  and  $\|P_{4|G} \otimes P_{2|H}\|_F^2 \leq k^2$ .

Therefore, we get the following upper and lower bound:

$$\begin{aligned}
\|A\|_* &\leq \|A_\perp\|_* + k \|\Delta\|_F, \\
\|B\|_* &\geq \|B_\perp\|_* - k^2 \|\Delta\|_F.
\end{aligned}$$

If we require that

$$\|A_\perp\|_* + k \|\Delta\|_F \leq \|B_\perp\|_* - k^2 \|\Delta\|_F,$$

then we will have  $\|A\|_* \leq \|B\|_*$ .

We can derive similar condition for the relationship  $\|A\|_* \leftrightarrow \|C\|_*$ . Let

$$\theta := \min\{\|B_\perp\|_* - \|A_\perp\|_*, \|C_\perp\|_* - \|A_\perp\|_*\}.$$

We thus obtain an upper bound on the allowed perturbation:

$$\|\Delta\|_F \leq \frac{\theta}{k^2 + k}. \quad (25)$$

## 12 Recovery Conditions for Latent Tree

When latent variables  $H$  and  $G$  are independent, we have that  $P_{HG} = P_H P_G^\top$ . In this case,

$$\begin{aligned} \|B_\perp\|_* &= \|(P_{3|G} \otimes P_{1|H})(\text{diag}(P_G) \otimes \text{diag}(P_H))(P_{4|G} \otimes P_{2|H})^\top\|_* \\ &= \|(P_{3|G} \text{diag}(P_G) P_{4|G}^\top) \otimes (P_{1|H} \text{diag}(P_H) P_{2|H}^\top)\|_* \\ &= \|P_{34} \otimes P_{12}\|_* \\ &\geq \|P_{34} \otimes P_{12}\|_F \end{aligned} \quad (26)$$

and

$$\begin{aligned} \|A_\perp\|_* &= \|(P_{2|H} \odot P_{1|H}) P_H P_G^\top (P_{4|G} \odot P_{3|G})^\top\|_* \\ &= \|P_{12}(\cdot) P_{34}(\cdot)^\top\|_* \\ &= \|P_{12}(\cdot) P_{34}(\cdot)^\top\|_F \\ &= \|P_{34} \otimes P_{12}\|_F \end{aligned} \quad (27)$$

and thus

$$\|A_\perp\|_* \leq \|B_\perp\|_*.$$

Suppose now that  $H$  and  $G$  are not independent and thus we have  $P_{HG} = P_H P_G^\top + \Delta$ . The goal is to characterize all  $\Delta$ s, such that  $\|A\|_* \leq \|B\|_*$  still holds for any quartet. From the above formulas it follows that the upper bound on  $\Delta$  depends only on pairwise marginal distributions.

Since the perturbed version of  $P_H P_G^\top$  remains a joint probability table, all entries of the perturbation matrix  $\Delta$  have to sum to 0, *i.e.*,  $\mathbf{1}^\top \Delta(\cdot) = 0$ . We further assume that each column sum and each row sum of  $\Delta$  is also equal to 0, *i.e.*,  $\mathbf{1}^\top \Delta = \mathbf{0}$  and  $\Delta \mathbf{1} = \mathbf{0}$ . In this case,  $\mathbf{1}^\top \Delta(\cdot) = 0$  is satisfied automatically.

The recovery conditions for latent trees can be derived in two steps. The first step is to provide recovery conditions for those quartet relations corresponding to a single edge  $H - G$  in the tree (Figure 7, left). In the second step we study quartet relations corresponding to paths  $H - M_1 - M_2 - \dots - M_l - G$  in the tree (Figure 7, right). We provide a condition under which the recovery condition of such quartets is reduced to the recovery condition on quartets from step 1. That is, we provide a condition under which the perturbation on the path is guaranteed to be smaller than the maximum allowed perturbation on an edge.

Let

$$\delta := \max_{H-G \text{ an edge}} \|\Delta_{HG}\|_F.$$

Our goal is to obtain conditions on  $\delta$ , under which recovery of any quartet relation is guaranteed.

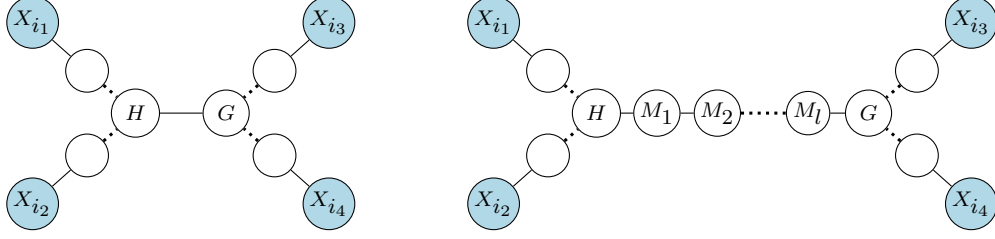


Figure 7: Topologies of quartets corresponding to a single edge  $H - G$  and to a path  $H - M_1 - M_2 - \dots - M_l - G$ .

### 12.1 Quartets Corresponding to a Single Edge

The first step is readily obtained from §11 if we assume that all CPTs (including  $P_{X_{i1}|H}$ ,  $P_{X_{i2}|H}$ ,  $P_{X_{i3}|G}$ ,  $P_{X_{i4}|G}$ ) have full rank. Let  $\theta_{\min} = \min_{\text{quarter } q} \theta_q$ . From (25), we have

$$\delta \leq \min \frac{\|B_{\perp}\|_* - \|A_{\perp}\|_*}{k^2 + k} = \frac{\theta_{\min}}{k^2 + k}. \quad (28)$$

### 12.2 Quartets Corresponding to a Path

**Path of independent latent variables.** For the second step, we start again from the fully factorized case (independent case). The joint probability table  $P_{HG}$  of the two end points in a path  $H - M_1 - M_2 - \dots - M_l - G$  is

$$\begin{aligned} P_{HG} &= P_{H|M_1} P_{M_1|M_2} \dots P_{M_l|G} P_G \\ &= P_{HM_1} \text{diag}(P_{M_1})^{-1} P_{M_1 M_2} \text{diag}(P_{M_2})^{-1} \dots \text{diag}(P_{M_l})^{-1} P_{M_l G} \\ &= P_H P_{M_1}^{\top} \text{diag}(P_{M_1})^{-1} P_{M_1} P_{M_2}^{\top} \text{diag}(P_{M_2})^{-1} \dots \text{diag}(P_{M_l})^{-1} P_{M_l} P_G^{\top} \\ &= P_H (P_{M_1}^{\top} \text{diag}(P_{M_1})^{-1}) P_{M_1} (P_{M_2}^{\top} \text{diag}(P_{M_2})^{-1}) \dots \text{diag}(P_{M_l})^{-1} P_{M_l} P_G^{\top} \\ &= P_H \mathbf{1}^{\top} P_{M_1} \mathbf{1}^{\top} \dots \mathbf{1}^{\top} P_{M_l} P_G^{\top} \\ &= P_H P_G^{\top}, \end{aligned}$$

where we have used  $P_{M_i}^{\top} \text{diag}(P_{M_i}(\cdot))^{-1} = \mathbf{1}^{\top}$ .

**Path of dependent latent variables.** Next, we add perturbation matrices to the joint probability tables associated with each edge  $M_i - M_j$  in the tree and assume that the resulting joint probability table  $P_{M_i M_j} = P_{M_i} P_{M_j}^{\top} + \Delta_{ij}$  has full rank. Furthermore, we assume that the resulting joint probability table  $P_{HG}$  of the two end points in a path  $H - M_1 - M_2 \dots M_l - G$  also

has full rank. We have

$$\begin{aligned}
P_{HG} &= P_{H|M_1} P_{M_1|M_2} \cdots P_{M_l|G} P_G \\
&= P_{H M_1} \text{diag}(P_{M_1})^{-1} P_{M_1 M_2} \text{diag}(P_{M_2})^{-1} \cdots \text{diag}(P_{M_l})^{-1} P_{M_l G} \\
&= (P_H P_{M_1}^\top + \Delta_1) \text{diag}(P_{M_1})^{-1} (P_{M_1} P_{M_2}^\top + \Delta_2) \text{diag}(P_{M_2})^{-1} \cdots \text{diag}(P_{M_l})^{-1} (P_{M_l} P_G^\top + \Delta_l) \\
&= P_H P_{M_1}^\top \text{diag}(P_{M_1})^{-1} P_{M_1} P_{M_2}^\top \text{diag}(P_{M_2})^{-1} \cdots \text{diag}(P_{M_l})^{-1} P_{M_l} P_G^\top \\
&\quad + 0 \text{ (terms not involving all the } \Delta \text{s will all be zero)} \\
&\quad + \Delta_1 \text{diag}(P_{M_1})^{-1} \Delta_2 \text{diag}(P_{M_2})^{-1} \cdots \text{diag}(P_{M_l})^{-1} \Delta_l \\
&= P_H P_G^\top + \Delta_1 \text{diag}(P_{M_1})^{-1} \Delta_2 \text{diag}(P_{M_2})^{-1} \cdots \text{diag}(P_{M_l})^{-1} \Delta_l.
\end{aligned} \tag{29}$$

The reason why we do not need to perturb the term  $\text{diag}(P_{M_i})^{-1}$  is that if  $\tilde{P}_{M_i}$  is the perturbed  $P_{M_i}$ ,

$$\tilde{P}_{M_i} = \tilde{P}_{M_i M_j} \mathbf{1} = (P_{M_i} P_{M_j}^\top + \Delta_{ij}) \mathbf{1} = P_{M_i} P_{M_j}^\top \mathbf{1} + \mathbf{0} = P_{M_i},$$

since  $\Delta_{ij} \mathbf{1} = \mathbf{0}$ . And the reason why terms not involving all the  $\Delta$ s will all be zero is that such terms contain either  $\mathbf{1}^\top \Delta = \mathbf{0}^\top$  or  $\Delta \mathbf{1} = \mathbf{0}$ .

Now, from (29) it follows that the perturbation corresponding to the path  $H - M_1 - M_2 - \cdots - M_l - G$  is

$$\Delta := \Delta_1 \text{diag}(P_{M_1})^{-1} \Delta_2 \text{diag}(P_{M_2})^{-1} \cdots \text{diag}(P_{M_l})^{-1} \Delta_l. \tag{30}$$

**Bounding the perturbation on the path.** We still need to show under which condition  $\Delta$  from (30) will satisfy  $\|\Delta\|_F \leq \delta$ . Assume that the smallest entry in a marginal distribution of an internal node is bounded from below by  $\gamma_{\min}$ , *i.e.*,

$$\gamma_{\min} := \min_{\text{hidden node } H} \min_i P_H(i).$$

Then we have

$$\begin{aligned}
\|\Delta\|_F &= \|\Delta_1 \text{diag}(P_{M_1})^{-1} \Delta_2 \text{diag}(P_{M_2})^{-1} \cdots \Delta_l\|_F \\
&\leq \|\Delta_1 \text{diag}(P_{M_1})^{-1}\|_F \|\Delta_2 \text{diag}(P_{M_2})^{-1}\|_F \cdots \|\Delta_l\|_F \\
&\leq \frac{\delta^l}{\gamma_{\min}^{l-1}}.
\end{aligned}$$

The perturbation  $\Delta$  on the path  $H - M_1 - M_2 \cdots M_l - G$  is bounded by  $\delta$  if  $\frac{\delta^l}{\gamma_{\min}^{l-1}} \leq \delta$ , *i.e.*, if

$$\delta \leq \gamma_{\min}. \tag{31}$$

From (28) and (31) we arrive at the condition for successful quartet test for all quartets

$$\delta \leq \min \left\{ \frac{\theta_{\min}}{k^2 + k}, \gamma_{\min} \right\}.$$

Intuitively, it means that the size of the perturbation  $\delta$  away from independence can not be too large. In particular, it has to be small compared to the smallest marginal probability  $\gamma_{\min}$  of a hidden state; it also has to be small compared to the smallest excessive dependence  $\theta_{\min}$ .

### 13 Statistical Guarantee for the Quartet Test

Based on the concentration result for nuclear norm in (11), we have that, given  $m$  samples, the probability that the finite sample nuclear norm deviates from its true quantity by  $\epsilon := \frac{2\sqrt{2\tau}}{\sqrt{m}}$  is bounded

$$\mathbb{P} \left\{ \|\hat{A}\|_* \geq \|A\|_* + \epsilon \right\} \leq 2e^{-\frac{m\epsilon^2}{8}} \quad \text{and} \quad \mathbb{P} \left\{ \|\hat{B}\|_* \leq \|B\|_* - \epsilon \right\} \leq 2e^{-\frac{m\epsilon^2}{8}}, \quad (32)$$

where we have used  $\tau = \frac{m\epsilon^2}{8}$ . Now we can derive the probability of making an error for individual quartet test. First, let  $q = \{\{i_1, i_2\}, \{i_3, i_4\}\}$  and

$$\alpha = \min \{ \|B(q)\|_* - \|A(q)\|_*, \|C(q)\|_* - \|A(q)\|_* \}.$$

Then, for sufficiently large  $m$ , we can bound the error probability by

$$\begin{aligned} & \mathbb{P} \{ \text{Quartet test returns incorrect result} \} \\ &= \mathbb{P} \left\{ \|\hat{A}\|_* \geq \|\hat{B}\|_* \text{ or } \|\hat{A}\|_* \geq \|\hat{C}\|_* \right\} \\ &\leq \mathbb{P} \left\{ \|\hat{A}\|_* \geq \|\hat{B}\|_* \right\} + \mathbb{P} \left\{ \|\hat{A}\|_* \geq \|\hat{C}\|_* \right\} \quad (\text{union bound}) \\ &= \mathbb{P} \left\{ \|\hat{A}\|_* - \|A\|_* + \|B\|_* - \|\hat{B}\|_* \geq \|B\|_* - \|A\|_* \right\} \\ &\quad + \mathbb{P} \left\{ \|\hat{A}\|_* - \|A\|_* + \|C\|_* - \|\hat{C}\|_* \geq \|C\|_* - \|A\|_* \right\} \\ &\leq \mathbb{P} \left\{ \|\hat{A}\|_* - \|A\|_* \geq \frac{\|B\|_* - \|A\|_*}{2} \right\} + \mathbb{P} \left\{ \|B\|_* - \|\hat{B}\|_* \geq \frac{\|B\|_* - \|A\|_*}{2} \right\} \\ &\quad + \mathbb{P} \left\{ \|\hat{A}\|_* - \|A\|_* \geq \frac{\|C\|_* - \|A\|_*}{2} \right\} + \mathbb{P} \left\{ \|C\|_* - \|\hat{C}\|_* \geq \frac{\|C\|_* - \|A\|_*}{2} \right\} \\ &\leq \mathbb{P} \left\{ \|\hat{A}\|_* - \|A\|_* \geq \frac{\alpha}{2} \right\} + \mathbb{P} \left\{ \|B\|_* - \|\hat{B}\|_* \geq \frac{\alpha}{2} \right\} \\ &\quad + \mathbb{P} \left\{ \|\hat{A}\|_* - \|A\|_* \geq \frac{\alpha}{2} \right\} + \mathbb{P} \left\{ \|C\|_* - \|\hat{C}\|_* \geq \frac{\alpha}{2} \right\} \\ &\leq 8e^{-\frac{m\alpha^2}{32}} \end{aligned}$$

### 14 Statistical Guarantee for the Tree Building Algorithm

Let  $\alpha_q = \min \{ \|B(q)\|_* - \|A(q)\|_*, \|C(q)\|_* - \|A(q)\|_* \}$ . We define

$$\alpha_{\min} = \min_{\text{quartet } q} \alpha_q.$$

For a latent tree with  $d$  observed variables, the tree building algorithm described in the paper requires  $O(d \log d)$  calls to the quartet test procedure. The probability that the tree is constructed incorrectly is bounded by the probability that either one of these quartet tests returns incorrect



result. That is

$$\begin{aligned}
& \mathbb{P}\{\text{The latent tree is constructed incorrectly}\} \\
& \leq \mathbb{P}\{\text{Either one of the } O(d \log d) \text{ quartet tests returns incorrect result}\} \\
& \leq c \cdot d \log d \cdot \mathbb{P}\{\text{quartet test returns incorrect result}\} \quad (\text{union bound}) \\
& \leq 8c \cdot d \log d \cdot e^{-\frac{m\alpha^2}{32}},
\end{aligned}$$

which implies that the probability of constructing the tree incorrectly decreases exponentially fast as we increase the number of samples  $m$ .